# R Code & Output

*Huafeng Zhang*

*4/21/2017*

## Data Manpulation

```
setwd("~/Desktop/BIGbanks")
Banks16 <- read.csv("Banks16.csv", header = TRUE)
Assets<- read.csv("Assets16.csv",header = TRUE)
Banks16<- left_join(Banks16, Assets, by="cert")

# Add latitude & longitude for the U.S. map later
library(zipcode)
# citation("zipcode")
library(dplyr)
# citation("dplyr")
data(zipcode)
z2 <- zipcode %>%
  mutate(zip=as.numeric(zip)) %>%
  select(zip, lat=latitude, long=longitude)
Banks16$zip<- as.numeric(Banks16$zip)
Banks16_geo_comp <- left_join(Banks16, z2, by="zip")
```

## Removed Banks with Missing Values

```
# 85 NAs
sum(is.na(Banks16))
```

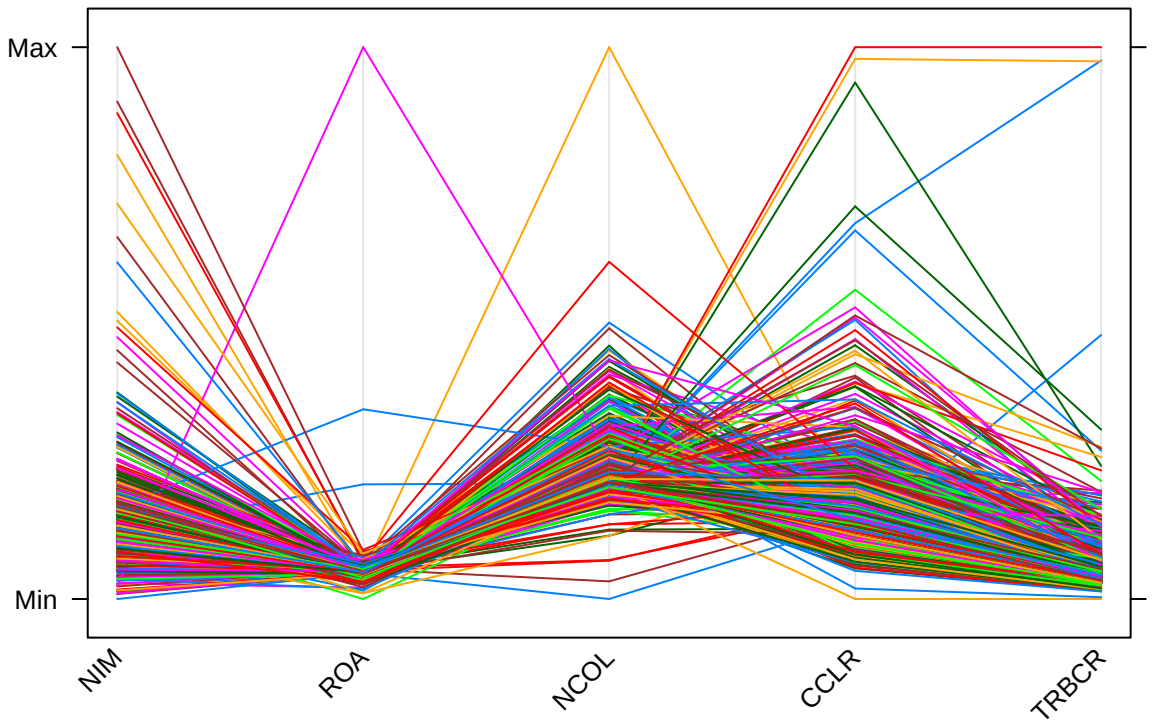
```
## [1] 85
```

```
Banks16_draft<-Banks16_geo_comp[complete.cases(Banks16_geo_comp), ]

# Make the names of the five varables consitent with the terms I used in my paper
names(Banks16_draft)[12:16]<- c("NIM","ROA","NCOL", "CCLR", "TRBCR")

# 48 banks was removed (5897-5849)
nrow(Banks16)-nrow(Banks16_draft)
```

```
## [1] 48
```

## PCP for the Data Set that is Used to Cluster (try to find potential skewed banks)

```
library(lattice)
# citation("lattice")
parallelplot(Banks16_draft[ ,12:16],scale=list(x=list(rot=45)),horizontal=FALSE)
```




### Reasoning for Excluding the Four Banks (P.S. the values of the assets are all in thousands (000's))

```
# Summary of all banks' assets
summary(Banks16_draft$asset_16)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 4.785e+03 9.759e+04 2.048e+05 2.855e+06 4.763e+05 2.083e+09
```

```
# The four banks' assets
TRBCR3<- head(sort(Banks16_draft$TRBCR, decreasing = TRUE),5 )
ROA1<-  head(sort(Banks16_draft$ROA, decreasing = TRUE),5 )
Bank4<- filter(Banks16_draft, Banks16_draft$cert %in% c(27575,18923, 21761,33831))
Assets16_Bank4<- Bank4$asset_16
Assets16_Bank4
```

```
## [1] 137852  60456  17419 216513
```

### Removed the Skewed Banks

```
Banks16_cont<- Banks16_draft[!Banks16_draft$cert %in% c(27575,18923, 21761,33831), ]
```

### Variable Summary for the Final Version Data Set

```
library(knitr)
Banks16_var<- Banks16_cont[ ,-c(1:11, 17:19)]
colnames(Banks16_var)<- c("NIM","ROA","NCOL", "CCLR", "TRBCR")
```

```
Variable<- c("NIM","ROA", "NCOL", "CCLR", "TRBCR" )
total<- rep(nrow(Banks16_cont),5)
n_NAs<- as.integer(c(10,9,48,9,9)) # From summary(Banks16_geo_comp)
Size_withoutNAs<- as.integer(total-n_NAs)
Mean_withNAs<- round(c(3.680,1.0312 ,0.16777,11.677,26.310),digits=3)
Mean_withoutNAs<- round(c(3.6983,0.9383 ,0.16767 ,11.274,18.840),digits = 3) # From summary(Banks16_var
SD1<- sqrt(var(Banks16_var$NIM))
SD2<- sqrt(var(Banks16_var$ROA))
SD3<- sqrt(var(Banks16_var$NCOL))
SD4<- sqrt(var(Banks16_var$CCLR))
SD5<- sqrt(var(Banks16_var$TRBCR))
SD_withoutNAs<- round(c(SD1,SD2,SD3,SD4,SD5),digits = 3)
var_sum<- data.frame(Variable,Size_withoutNAs,n_NAs,Mean_withNAs,Mean_withoutNAs,SD_withoutNAs)
knitr::kable(var_sum)
```

| Variable | Size_withoutNAs | n_NAs | Mean_withNAs | Mean_withoutNAs | SD_withoutNAs |
|---|---|---|---|---|---|
| NIM | 5860 | 10 | 3.680 | 3.698 | 1.086 |
| ROA | 5861 | 9 | 1.031 | 0.938 | 1.278 |
| NCOL | 5822 | 48 | 0.168 | 0.168 | 0.536 |
| CCLR | 5861 | 9 | 11.677 | 11.274 | 3.888 |
| TRBCR | 5861 | 9 | 26.310 | 18.840 | 10.019 |

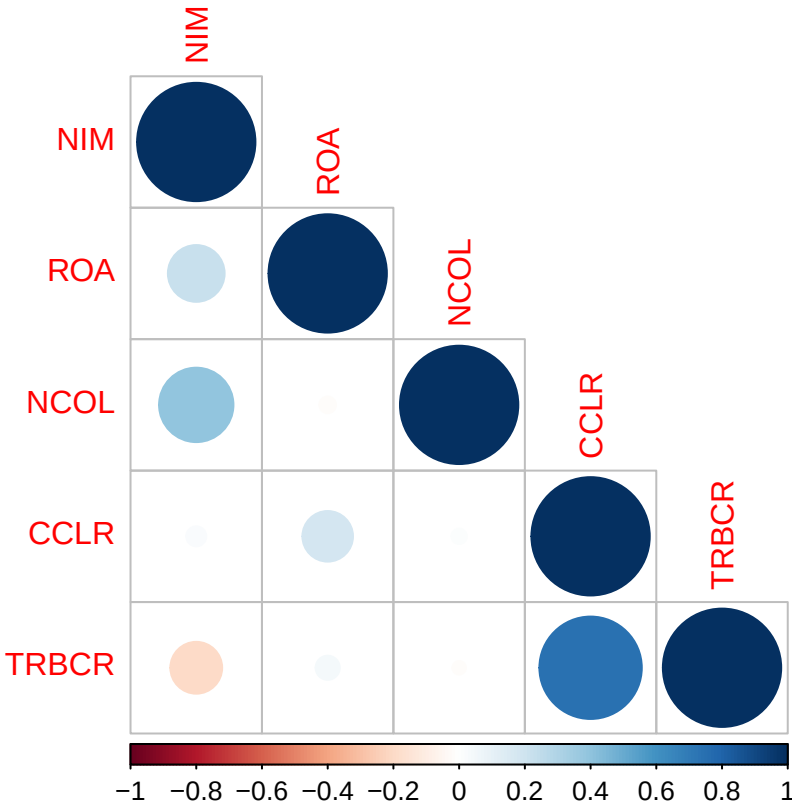
## Check Correlation between Variables

```
library(corrplot)
# citation("corrplot")
cor_banks<- cor(scale(Banks16_var[ ,1:5]))
knitr::kable(cor_banks, caption="Correlation Between Variables in the Study")
```

Table 2: Correlation Between Variables in the Study

| | NIM | ROA | NCOL | CCLR | TRBCR |
|---|---|---|---|---|---|
| NIM | 1.0000000 | 0.2298826 | 0.3952213 | 0.0283956 | -0.1916184 |
| ROA | 0.2298826 | 1.0000000 | -0.0196332 | 0.1825480 | 0.0414919 |
| NCOL | 0.3952213 | -0.0196332 | 1.0000000 | 0.0167341 | -0.0129090 |
| CCLR | 0.0283956 | 0.1825480 | 0.0167341 | 1.0000000 | 0.7487761 |
| TRBCR | -0.1916184 | 0.0414919 | -0.0129090 | 0.7487761 | 1.0000000 |

```
corrplot(cor_banks,type="lower")
```

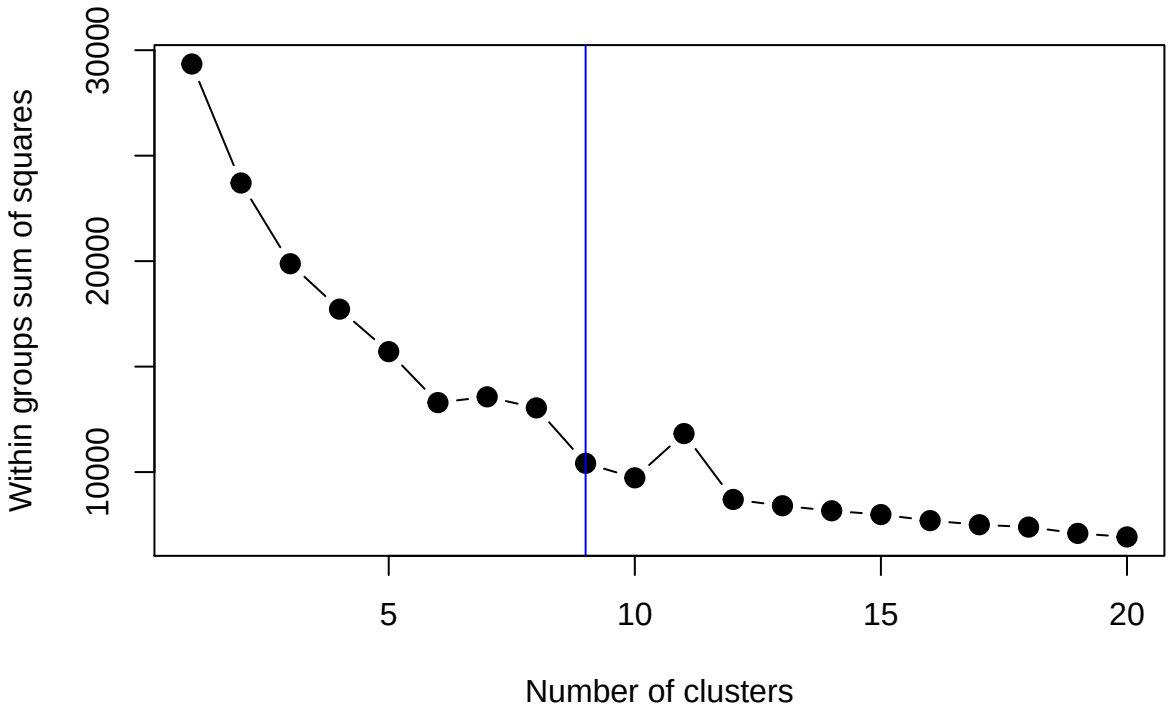



### Scale the Five Variables

```
scale_var<- scale(Banks16_var,center = TRUE, scale = TRUE)
```

### Find the Optimal NUmber of Clusters (the Elbow Method)

```
set.seed(66666)
c_df<- data.frame()
for (k in 1:20){
    c_w<- kmeans(scale_var,centers=k,iter.max = 100)
   c_df<- rbind(c_df,cbind(k,c_w$tot.withinss))
  }
names(c_df)<- c("Clutser","Total within SS")
plot(c_df,type = "b",xlab="Number of clusters",ylab="Within groups sum of squares",pch=20,cex=2)
abline(v=9,col="blue")
```




### Run Clustering and Find the Cluster Summary

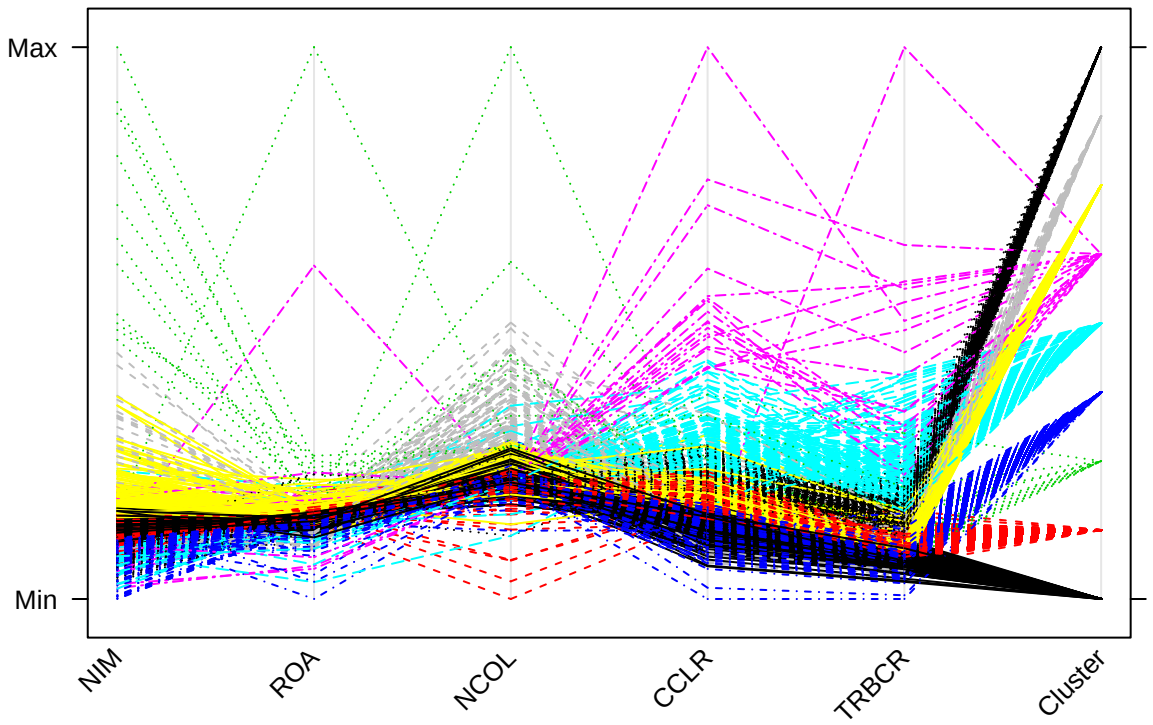
```
k9<- kmeans(scale_var,9)
Banks16_cont$Cluster<- k9$cluster
Cluster_means<- Banks16_cont %>%
                group_by(Cluster) %>%
                summarise(mean(NIM),mean(ROA),mean(NCOL),mean(CCLR),mean(TRBCR))
Cluster_means$ Size<- c(k9$size)
knitr::kable(Cluster_means,caption="Cluster Summary")
```

Table 3: Cluster Summary

| Cluster | mean(NIM) | mean(ROA) | mean(NCOL) | mean(CCLR) | mean(TRBCR) | Size |
|---|---|---|---|---|---|---|
| 1 | 3.870394 | 0.9344076 | 0.1199350 | 9.483252 | 13.70153 | 2075 |
| 2 | 3.690945 | 1.1114096 | 0.0496430 | 12.358850 | 19.31060 | 1143 |
| 3 | 18.277745 | 10.2926731 | 4.3478333 | 18.618990 | 26.59167 | 12 |
| 4 | 2.957853 | 0.5178841 | 0.1139025 | 9.336644 | 17.01699 | 1224 |
| 5 | 3.010524 | 0.6005046 | 0.1124774 | 22.286575 | 52.81566 | 161 |
| 6 | 3.162718 | 2.9453427 | 0.0547126 | 45.728966 | 97.97948 | 19 |
| 7 | 4.953804 | 1.6089130 | 0.2821769 | 11.334864 | 16.17321 | 587 |
| 8 | 4.869872 | 0.0855901 | 3.1199110 | 11.378513 | 18.24960 | 72 |
| 9 | 3.147923 | 0.7586064 | 0.1345359 | 15.242313 | 30.79582 | 577 |

### PCP Sorted by Clusters

```
Banks16_var$Cluster<- k9$cluster
parallelplot(Banks16_var,col=Banks16_cont$Cluster,lty=Banks16_cont$Cluster,scale=list(x=list(rot=45)),h
```



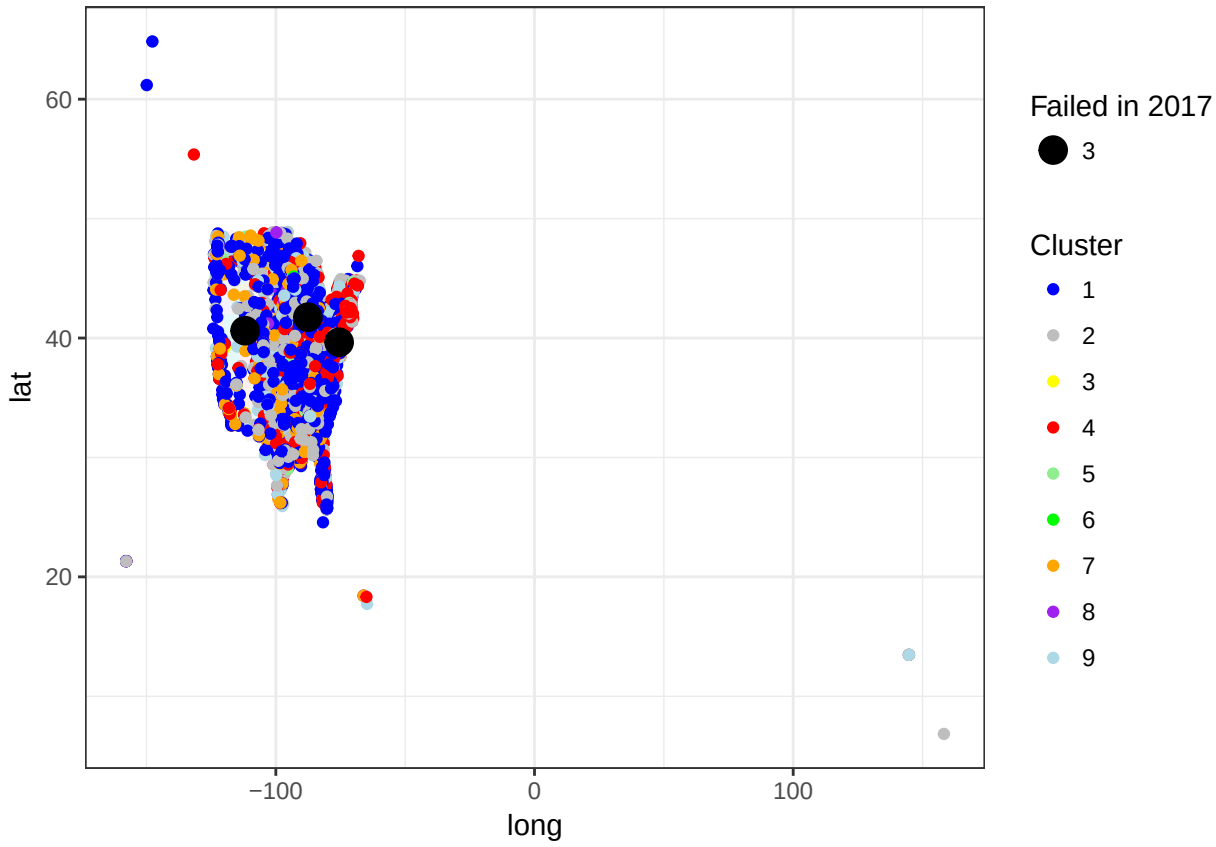

### Banks Failed in 2017

```
# Data manipulation
Failed <- read.csv("Failed.csv", header = TRUE)
Failed$Closing.Date <- as.character(Failed$Closing.Date)
Failed$Closing.Date <- as.character(Failed$Closing.Date)
substrRight <- function(x,n){
  substr(x, nchar(x)-n+1,nchar(x))
}
Failed$Year <- as.factor(substrRight(Failed$Closing.Date, 2))
colnames(Failed)[4]<- "cert"
fb <- select(Failed, cert, fYear=Year)
Banks16_cont_f<- left_join(Banks16_cont, fb, by="cert")

# Find banks failed in 2017
Banks16_cont_f17<- subset(Banks16_cont_f, fYear=="17")
View(Banks16_cont_f17[ ,-c(2:12)])

# Separate the origin clustering data set by using Banks failed in 2017 or not
Banks16_cont_Notf17<- subset(Banks16_cont_f, fYear=NA)
```

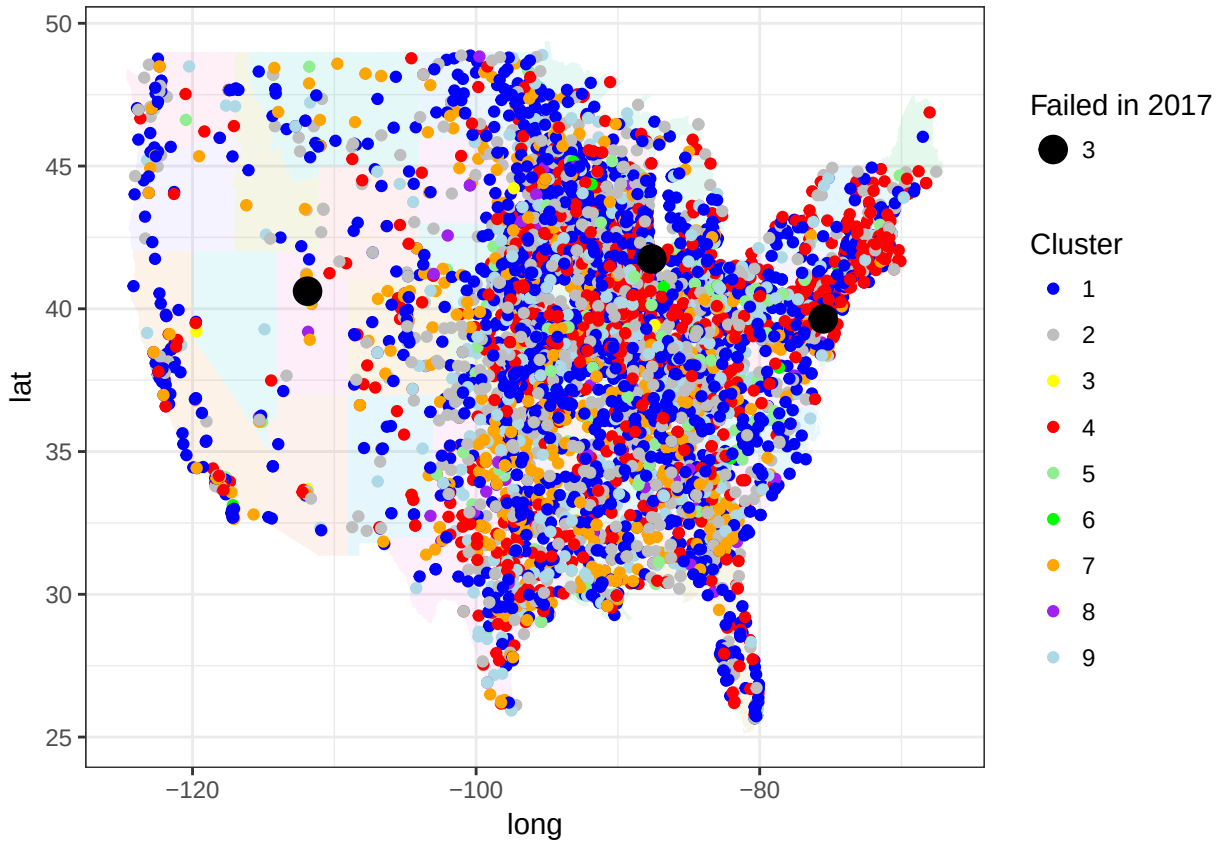
### Visulize Banks in the U.S. Map

```
library(maps)
library(plotly)
# citation("maps")
# citation("plotly")
```

```
states <- map_data("state")
Banks16_cont_Notf17$Cluster<- as.factor(Banks16_cont_Notf17$Cluster)
Banks16_cont_f17$Cluster<- as.factor(Banks16_cont_f17$Cluster)
map1 <- ggplot() +
  geom_polygon(data = states, aes(x=long, y=lat, fill=region, group=group), alpha=.1) +
  geom_point(data =Banks16_cont_Notf17, aes(x=long, y=lat, color=Cluster)) +
   geom_point(data =Banks16_cont_f17, aes(x=long, y=lat,size=3)) +
  geom_text(label=Banks16_cont_f$Cluster,hjust=0, vjust=0)+ scale_color_manual(values=c("blue", "grey",
map1
```




```
# ggplotly(map1)

## In order to see the banks that failed in 2017 in a more clearer way
Banks16_geo_cont_Notf17 <- filter(Banks16_cont_Notf17, long < -67 & long > -130  & lat>25 )
map2 <- ggplot() +
  geom_polygon(data = states, aes(x=long, y=lat, fill=region, group=group), alpha=.1) +
  geom_point(data =Banks16_geo_cont_Notf17, aes(x=long, y=lat, color=Cluster)) +
   geom_point(data =Banks16_cont_f17, aes(x=long, y=lat,size=3)) +
  geom_text(label=Banks16_cont_f$Cluster,hjust=0, vjust=0)+ scale_color_manual(values=c("blue", "grey",
map2
```




```
# ggplotly(map2)
```